SEEDS

Working Paper Three:

Data Service Provider Model,

Functional Areas

April 24, 2002

G. Hunolt, SGT, Inc.

Outline

1 Introduction

This working paper is the third of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation team in estimating the life cycle costs of future ESE data service providers and supporting systems, where 'data service provider' is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the project, obtaining feedback and guidance from ESDIS (Earth Science Data and Information System project) SOO (Science Operations Office) and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each paper that appears represents a snapshot in time, with the work in various stages of completion; as work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

This third working paper of the set describes the general data service provider reference model developed for the LOS/CE study, and discusses the functional areas included in the model.

The functional area descriptions in this paper reflect results of the February, 2002, SEEDS Community Workshop, comments and recommendations made at the workshop and in white papers submitted to the workshop.

2 Data Service Provider Reference Model

This section describes the Data Service Provider (DSP) Reference Model, a functional model of a generic data service provider.

The reference model has three related aspects:

1) A set of 'functional areas' that collectively comprise the full range of functions that a data service provider might perform and the areas of cost that must be considered by the cost estimation by analogy model.

2) A set of parameters for each functional area that constitute a quantitative description of the workload, staff effort, and any other factors that contribute to cost for that area, additional 'roll-up' parameters that sum items such as staff effort across the functional areas, and other parameters like labor rates that are required for cost estimation.

3) A set of requirements and levels of service for each functional area.

These three aspects of the model are closely coupled to ensure the internal consistency of the model. The set of functional areas is the underpinning; both the model parameters and requirements / levels of service are organized according to the functional areas. The requirements / levels of service and the model parameters are coupled in that the definitions of the requirements / levels of service embody model parameters. This integration of the three aspects of the model is intended to ensure that estimated costs are driven by and traceable to requirements to the fullest extent possible.

The intent of the descriptions of the functional areas (see Section 3 below) and the corresponding requirements / levels of service (see Working Paper 5, "General Data Service Provider Reference Model - Requirements / Levels of Service") is to provide a reasonably full description of the abstract ESE data service provider, and to reflect the concerns expressed in the February, 2002, SEEDS community workshop. The ability of the cost estimation by analogy approach to reflect the full range of detail described in the functional areas and requirements / levels of service will be limited by the information available in the comparables database and the feasibility of reasonable assumptions where information is not available. This will be reflected in the reference model's parameter set.

2.1 Functional Areas - Areas of Cost

The functional areas of the reference model are defined in Section 3 of this paper. Some of the areas are not strictly speaking "functional" in nature (such as 'facility / infrastructure') but are needed to ensure that all significant cost areas are included.

The functions / areas of cost span the full life cycle from implementation through operations. Implementation includes capital and staff costs associated with developing, implementing, integrating and testing the data service provider's data and information system, and facility start-up / preparation costs. Implementation is assumed to be spread over a specified number of years. Implementation can overlap the start of operations. Implementation can also recur during the operating period, allowing for system expansion, enhancement, or replacement, i.e. 'technology refresh'. Operations includes hardware maintenance, sustaining engineering, operations staff, supplies (e.g. storage and archive media), recurring facility costs, etc., for the expected lifetime of the activity.

2.2 Reference Model Parameters

The parameters of the reference model are defined in detail in Working Paper 4, "General Data Service Provider Reference Model - Model Parameters".

The scope of the parameters spans implementation and operations, year by year over the specified lifecycle of the data service provider, and includes cost elements as well as workload factors and high level system configuration information.

The implementation and operations parameters will be broken down into outputs to be provided by the model, internal (derived) parameters used by the model, and inputs required by the model.

The cost estimation relationships to be used by the model will be derived from information describing actual date centers or other data service providers comparable to future ESE data service providers. Raw information received from the data service providers will be mapped to the standard reference model parameter set to build the model's comparables database, so that the database will contain an internally consistent set of parameters.

The comparables database will be used to derive the cost estimating relationships (CERs) that allow estimation of the outputs given the inputs for independent cases. This will include testing the model against independent data for an actual data service provider (for whom the actual outputs are known) and eventual use of the model to estimate the costs for a putative new ESE data service provider.

2.3 Reference Model Requirements / Levels of Service

The requirements / levels of service of the reference model are presented in Working Paper 5, "General Data Service Provider Reference Model - Requirements / Levels of Service".

The general data service provider reference model will map to a general requirements template, a statement of requirements / levels of service for a generic data service provider, in which the requirements / levels of service are defined for all of the functional areas included in the model.

The requirements / levels of service are a template in that they contain placeholders for quantitative parameters that will be defined for a specific instance of a data service provider. For example, a requirement in the template might be that "the data service provider shall provide an archive capacity of [number TB]". A data service provider of a type that would include providing an archive would have that item in its template. If the mission of the data service provider required that it archive certain data streams and generated products that would accumulate to a total volume of 100 TB, then that value would be inserted into the template, with the result being a specific requirement for that data service provider (i.e., "the data service provider shall provide an archive capacity of 100 TB") that could then be used in the process of generating a cost estimate for the data service provider.

2.4 Reference Model Subsets - Logical Data Service Provider Types

The general data service provider reference model includes all functions / areas of cost that a generic data service provider might perform. While an actual working data service provider could conceivably perform all of the functions included in the model, most if not all actual data service providers perform a subset of them, e.g. most providers will not have a requirement in the area of instrument / mission operations. Many well known actual data centers such as the NASA Distributed Active Archive Centers (DAACs) or the NOAA national data centers perform a subset of the general set of functions. Some data service providers, e.g. MODAPS (the MODIS Adaptive Processing System) as a sample of a science team processing facility that does not perform archive or general user distribution), are different in function from many well known data centers but fit within the framework of the data service provider reference model.

The Cost Estimation Tool will allow a planner (for example) planning a data service to support a flight project, to:

1. select those functions that are required for his/her particular mission (in effect to create a 'custom' subset of the general model);
2. specify the particular mission requirements the real instantiation of it must meet (e.g. data volumes to be ingested, processed, stored, and/or distributed);
3. produce an estimated cost for implementing and operating it.

A set of ‘logical data service provider types’ has been defined to enable overall ESE data service architecture studies (where a ‘data service architecture’ is a collection of data service providers and the interconnections between them), and as an option available for use by planners of individual data service provider activities. Each of type is a functional subset of the general reference model organized around a defined class of ESE role or mission. These are ‘logical’ types in that there is no explicit or implicit 1:1 mapping of an instance of a logical data service provider type to a physical entity. While some actual data service providers might match a logical type, most will perform the functions of more than one logical type, and may also perform multiple data service activities within the scope of a type (such as a DAAC that performs archive and distribution for several flight projects). Because the logical data service provider types are only a few of the possible subsets of the general model, they constitute an open set to which additions (and subtractions) can be readily made as needed to facilitate architecture trade studies or other uses.

The current set of logical data service provider types is described in Working Paper 6, “ESE Logical Data Service Provider Types”, which describes each type and indicates the subset of the functional areas and requirements / levels of service that apply to it.

3 General DSP Reference Model - Functional Areas

This section describes the functional areas / areas of cost that comprise the Data Service Provider Reference Model. They describe the full range of functions of an abstract general data service provider. It is unlikely that an actual ESE data service provider would perform in all of the functional areas; different ones would perform in different subsets of the full set, and would perform at different levels (i.e. provide different levels of service) within functional areas.

The functional areas are primarily focused on operating activities of the data service provider. The data service provider also has additional responsibilities that require high levels of expertise in science in the discipline(s) supported by the provider, data management expertise, and information technology expertise. The Management area (see Section 3.13 below) includes lead, site-level responsibilities in these areas, and the Technical Coordination area (see Section 3.11 below) includes coordination with other ESE data service providers and broader communities in these areas.

The intent is not to provide exhaustive descriptions in great detail of every possible aspect of each of the functional areas, but rather to describe key aspects of each that are of greatest concern to either users or data service provider operators or planners or significant cost drivers.

The following sections present working definitions of the functional areas that make up the data service provider reference model.

3.1 Ingest

The ingest functional area includes receiving, reading, quality checking, cataloging, of incoming data (including metadata, documentation, etc.) to the point of insertion into the archive. Ingest can be manual or electronic with manual steps involved in quality checking, etc.

Incoming data can be received from external sources or internally generated. Ingest can include format conversion, metadata extraction, or other preparation of incoming data for archive or use within the data service provider. Ingest includes verifying that all data made available for ingest has been successfully ingested, with exceptions tracked and accounted for. Ingest must be accomplished in a timely manner as needed to meet mission requirements of the data services provider.

3.2 Processing

The processing functional area includes the generation and quality checking of new derived data products from data or products that have been ingested, or previously generated, generally on a routine, operational basis. Operational processing can be on demand as well as scheduled. Operationally generated products are often ‘standard products’ characterized by a peer reviewed, validated, reasonably stable, ‘science quality’ processing algorithm.

Processing includes ad hoc, non-operational generation of products that can include responding to requests for data mining or generation of special subsets. Processing includes process control (production planning, scheduling, monitoring, etc.) as well as product generation per se. Processing also includes reprocessing of new versions of previously generated products, either according to a reprocessing schedule or plan, or as allowed within a specified overall reprocessing capacity.

Where science or applications needs require simultaneous measurements from multiple instruments, processing performed by a data service provider can include data integration - mapping parameters from different sources to a common spatial / temporal base.

Processing can also include ‘data mining’, where software may search through many of the holdings of a data service provider for items meeting certain criteria.

The data service provider may receive the software that embodies product generation algorithms from outside developers (e.g. some Terra instrument teams for the DAACs currently) who are responsible for the initial delivery and for delivering updated versions. Where quality, especially science quality, of products remains the responsibility of an outside developer, processing includes supporting quality checking by the science software developer. Support provided by the data service provider for integration and test of this 'science software' is included as an activity under processing. In cases where a data service provider develops algorithm software, that effort (i.e. development, integration, and test) is included under Implementation.

The data service provider may also accept software from science or applications users to produce a research product, perform data integration, or perform data mining.

3.3 Documentation

The documentation functional area includes the development (or upgrading of received) data and product documentation (including user guides, catalog interfaces, etc.) to meet SEEDS adopted documentation standards, including catalog information (metadata), user guides, etc., through consultation with data providers, algorithm developers, flight projects, etc. Knowledge capture is a critical concern - the data service provider must be committed to pro-actively capture knowledge of instruments, calibration, processing history, etc., from its data sources (e.g. instrument teams).

SEEDS adopted documentation standards may include FGDC (Federal Geographic Data Committee) metadata standards, documentation standards for long term archiving, Algorithm Theoretical Basis Documents (or equivalent, which must reflect 'as-built' algorithms), Data Software Interface Specifications, etc. When science needs require that multiple versions of a product be held, the documentation of each version must include the provenance information (e.g. processing algorithm) peculiar to it.

Documentation should include comments received from users on their experience with the data and products (product accuracy, usability, etc.), perhaps in the form of FAQ's (Frequently Asked Questions) for products, both from scientists on staff or working closely with the provider and from the general user community.

Documentation should include read software and other appropriate tools for data access kept current with commonly available technology. Documentation includes maintenance and refresh according to best industry practice or SEEDS policy.

Documentation needs will evolve, e.g. information relevant to intellectual property rights may be needed.

3.4 Archive

The archive functional area includes the insertion of data into archive storage, and data stewardship - management, handling and preservation of data, metadata, and documentation within a data service provider's archive. Inserted data can include data ingested from sources external to the site, or data/products generated on-site.

Data stewardship / preservation includes quality screening of data entering and exiting the archive, quality screening of archive media, tested and verified backup and restoration capability, and accomplishing migrations from one type of media to another.

Insertion into the archive can be electronic or manual (e.g. hanging tapes on a rack or popping them into a robotic silo).

3.5 Search and Order

The search and order functional area includes providing access to catalog information (a range of descriptive information to aid in selecting data and products) and a search and order capability to users, and receiving user requests for data.

"Search and order" in this context is used in a very broad sense; search and order includes support for system to system interactions as well as conventional search and order by users directly. For example, system to system interactions might include a program running on a user platform accessing the data service provider system directly, locating a needed product, and executing a protocol (e.g. for user registration, security) to gain access to it.

"Search", whether by a user directly or through a system-system interaction, implies applying criteria that might include geophysical parameter(s), spatial-temporal coverage, specific product names, etc., to the metadata describing available data and products and returning to the user listings supplemented by descriptive information of those data or product types and instances that meet the criteria.

"Order" implies a request/permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a selected subset of the results of) a search, is processed and accepted or denied.

Search and order can include providing local user interface and capability and/or providing an interface to a broader based, cross-site search and order capability (e.g. DAACs supporting search and order via the EOS Data Gateway).

3.6 Access and Distribution

The access and distribution functional area includes fetching the requested data from the archive, performing any subsetting, resampling, reformatting / format conversion (e.g. to a GIS (Geographic Information System) format), reprojection, or packaging, and providing the end product to the user by electronic means or on physical media.

"Access" is included to embrace a service allowing a program running on a user platform to access data and products from the data service provider directly, through an appropriate protocol, perhaps as a seamless extension of the system to system search and order described above.

Access and distribution can be performed on an operational basis, meaning in part that a data service provider will formally commit to terms of service in a level of service agreement or equivalent.

Access and distribution is an area likely to see substantial evolution in the next five to ten years, perhaps especially if distributed computing comes into play on a significant scale. Highly automated access techniques, software agents, and new tools for data discovery, access, integration from multiple sources, etc., will become available.

Note: Success from a user point of view may be even more dependent on a product's format than the speed of its delivery (what if a product is delivered in 30 seconds but the form is such that a user needs to spend several hours to be able to use it, vs a product delivered in 30 minutes but in a form that can be used directly?). The data service provider should take care to offer formats (whether as a default or an option) that are directly useable by the largest possible fraction of its user community.

3.7 User Support

The user support functional area includes support provided in direct contact with users by user support staff, including responding to queries, taking of orders, staffing a help desk (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.), etc. User support staff includes science expertise to assist users in selecting and using data and products.

The demands on user support will increase with the proliferation of data types, data sources, and tools for users, continuing or increasing the need for highly trained user support staff even as user interactions become more automated and more automated user support aids become available (beginning with on-line documentation, FAQ, etc.).

User support also includes outreach to potential new users and education / training for current or potential new users.

User support should also be a channel for feedback from the users to the data service provider, whether comments on particular data or products or on the provider's services and support.

User support includes coordination of user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program - see Technical Coordination.

3.8 Instrument / Mission Operations

The instrument / mission operations functional area includes monitoring instrument and spacecraft performance, generating instrument and spacecraft commands, and event scheduling (using NASA or other appropriate operational mission management services).

3.9 Sustaining Engineering

Sustaining engineering includes maintenance and enhancement of custom applications software (including any science software embodying processing algorithms developed by the site).

3.10 Engineering Support

Engineering support includes some or all of the following as applicable at a particular site: systems engineering, test engineering, configuration management, coordination of hardware maintenance by vendors, COTS procurement, installation of COTS upgrades, system administration, database administration, network/communications engineering, and security.

Engineering support is internal, directed toward the internal operation of the data service provider.

3.11 Technical Coordination

Technical coordination includes participation in SEEDS system level processes, including coordination on data management, data stewardship (including standards for content of life cycle data management plans), standards and best practices (including quality assurance standards and practices), interfaces, common metrics, and interoperability (e.g. for data access and integration), across / within SEEDS and with other systems and networks as needed to support the ESE program.

This area includes coordination on evolution of the overall ESE data service architecture (including an examination of the changing needs of the ESE science and applications program and the consequent impacts on the roles, missions, and services of ESE data service providers).

Technical coordination includes participation in SEEDS system level processes to coordinate user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program.

Technical coordination includes participation in SEEDS level and/or bilateral processes to coordinate production and delivery of products between ESE data service providers.

Technical coordination includes cooperating with other ESE data service providers in representing ESE / SEEDS in broader community processes in areas such as standards, interoperability, data management, security, etc.

Technical coordination, which by its nature includes engineering, is directed outward, supporting the data service provider as one element of a system of cooperating centers.

3.12 Implementation

Implementation includes development of, and making operational, the data and information system capabilities required by the data service provider to perform its mission, including design and implementation of the data system (hardware and system software) and applications software. Implementation can recur during the operating period as systems are expanded or replaced.

In addition to a major implementation effort, implementation can include ongoing applications software development. Implementation can include development of software tools for use by users to unpack, subset, or otherwise manipulate products provided by the data service provider.

In some cases applications software will include product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need. Applications software can include software to perform a 'data mining' or data integration operation to meet a user need.

3.13 Management

Management includes management and administration at the data service provider level ("front office") and direct management of functional areas. Management also includes staff with overall responsibility for internal and external science activities, information technology planning, and data stewardship.

Management includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology.

Management includes developing data stewardship practices, performing data administration with science advice (via the User Advisory Group and other appropriate bodies), developing and maintaining life cycle data management plans (which address data migrations).

Management also includes coordinating the science activities within the data service provider and its interaction with the ESE and broader science community, including a visiting scientist program, collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group (which includes representation from the science, applications, education, etc., communities as appropriate for a given data services provider) and any other ESE or broader advisory activities that may be appropriate.

Management also includes participation in SEEDS management processes, strategic planning, coordination with other data centers and activities beyond ESE/SEEDS.

Management includes performing supervisory, financial administration, and other administrative functions.

3.14 Facility / Infrastructure

Facility / Infrastructure includes provision and maintenance of a fully furnished and equipped, environmentally controlled, physically secure facility to house data service provider staff, systems, and data and information holdings, including a backup facility for its data and information holdings. An off-site backup facility would be one sufficiently removed from the data service provider's primary site such that a fire, tornado, or other event that destroys the primary site would be very to extremely unlikely to also destroy the backup site (a risk analysis would be performed on a site by site basis).

This area includes resource planning, logistics, supplies inventory and acquisition, and facility management.

This area includes maintenance of system and site security according to established NASA security policies and practices.

Facility / Infrastructure also includes a variety of non-staff cost factors such as supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

References and Acronyms

The References Section and the Acronym List for all of these Working Papers is in the document "References and Acronyms for the Levels of Service / Cost Estimation Working Papers ".